Report by: Anand Prashar

NLP Homework 5 – q3

Q:

Explore similarities in nytcounts.4k, which contains context counts for about 4000 words in a sample of New York Times. The news data was lowercased and URLs were removed. The context counts are for the 2000 most common words in twitter, as well as the most common 2000 words in the New York Times. (But all context counts are from New York Times.) The context counts only contain contexts that appeared for more than one word. The _le vocab contains the list of all terms in this data, along with their total frequency. **Choose at least six words. For each, show the output of show nearest() and comment on whether the output makes sense. Comment on whether this approach to distributional similarity makes more or less sense for certain terms.** Four of your words should be:

_ a name (for example: person, organization, or location)
_ a common noun
_ an adjective
_ a verb

You may also want to try exploring further words that are returned from a most-similar list from one of these. You can think of this as traversing the similarity graph among words. Complete the code used to generate answers for this question in q3.py and upload it to Vocareum for documentation purposes; provide written responses on CrowdMark below.

A:

```
experiment_list = ['edward','school','red','saved', 'eyebrows','church']

Name => 'edward'
Common noun => 'school', 'eyebrows'
Adjective => 'red'
Verb => 'saved'

Exploring further word returned from most-similar list to 'school'=> 'church'
```

Result of show nearest on above 6 words:

```
file nytcounts.4k has contexts for 3648 words

Experiment = edward
1: richard (0.865773624009)
2: robert (0.860447306541)
3: andrew (0.853929024452)
4: joseph (0.848822376331)
5: william (0.839134959519)
6: stephen (0.834130613871)
7: john (0.819646643676)
8: david (0.816140938498)
9: daniel (0.810912262385)
10: peter (0.809679325607)
```

Comment:
This is a Name.
With understanding that context count dimensions of vectors are words, before or after which, our experiment word occurred;
It makes sense that other nouns could have been associated to same word features; with their frequency showing how similar they were as Edward's if they were to replace it.

```
Experiment = school
1: schools (0.741096505683)
2: college (0.716161495973)
3: line (0.694353893027)
4: church (0.692936169261)
5: practice (0.692639540488)
6: experience (0.68964261712)
7: location (0.686896635079)
8: scenes (0.684265543442)
9: standards (0.68314025111)
10: movement (0.682270236675)
```

Comment:
This is a Common Noun.
Take an example comparison between 'school' and 'college'.
Some of common context word features are: player, players, Kansas … …
It makes sense that 'school' and 'college' could be interchangeably used as adjacent word to these features.

Experiment = **red**
1: japanese (0.774842102043)
2: french (0.773843527982)
3: russian (0.771601727641)
4: british (0.769760881161)
5: german (0.76308587011)
6: chinese (0.753910259605)
7: u.s. (0.753802151553)
8: senate (0.746977104327)
9: final (0.742776789257)
10: second (0.741986898927)

Comment:
This is an adjective.
It has some interesting observation. There is some property/feature set that heavily influences the reason of higher similarity.
For instance, 'red' and 'french' show similarity… and this relation is heavily influenced by the fact that the French drink/make **wine** and **wine** is red. **Wine** word appears in many forms and weights in both vectors.

Experiment = **saved**
1: accepted (0.819118221409)
2: rejected (0.799582216029)
3: reached (0.766923280243)
4: kept (0.76670376084)
5: made (0.759091263507)
6: raised (0.756470951843)
7: watched (0.742826298461)
8: blocked (0.742740943833)
9: pushed (0.74185754295)
10: lost (0.740670135264)

Comment:
This is a verb. On a glance, it does not make sense to see the lists as similar. Unfortunately, the result we got was heavily influenced by a feature word '**the**' which had a high weight.

Experiment = **eyebrows**
1: teeth (0.696998316688)
2: shoes (0.696253261621)
3: friends (0.687035537156)
4: hair (0.68532575405)
5: clothes (0.681900723376)

6: goals (0.678553582538)
7: skills (0.678436322527)
8: legs (0.670437263485)
9: classes (0.666140092645)
10: shots (0.662747196322)

Comment:
This is a Noun. Most of its common features come from words like-

"'s _", '_ ,', ', _', 'your _', "'' _", '_ .', 'and _', '. _', '_ and', '_ but', '_ are', 'her _', '_ among', 'the _', 'black _', '_ by', '_ with', 'those _', '_ just', 'my _', '_ because', '_ have', '_ for', '_ of', 'their _', '_ is', '_ in', 'some _', '_ at', '_ that', 'his _', '_ or'

It does make sense to substitute them from list, but does not make sense to consider both words similar.

Experiment = **church**
1: club (0.911603076953)
2: foundation (0.911124136997)
3: store (0.903660430834)
4: hospital (0.901291429841)
5: project (0.899371781413)
6: case (0.898702893185)
7: streets (0.89820350555)
8: trial (0.897505556383)
9: program (0.896075133092)
10: gallery (0.894118952333)

Comment:
This is a Noun. Exploring further word returned from most-similar list(taken from previous experiment 'school'). In same context, we can say this experiment captures that that context correlation with this word. There is no exact synonym similarity, but it gives an idea why these words could be used together/as replacement in same context.
Also, notice 'church' was a similar word to 'school', but reverse it not true. This word has other more similar words than 'school'

In summary:

For most types of Nouns and adjectives, it make more sense as context similarity are captured well in Sparse-Vector representation with Left and Right words and we see a cascading hierarchy. But in others like Verbs, it seems to lack the connection due to distraction in order to consider them as flowing hierarchy.